

---

# Adaptive Semisupervised Inference

---

**Martin Azizyan**

Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA, 15213-3890 USA

**Aarti Singh**

Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA, 15213-3890 USA

**Larry Wasserman**

Department of Statistics and  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA, 15213-3890 USA

## Abstract

Semisupervised methods inevitably invoke some assumption that links the marginal distribution of the features to the regression function of the label. Most commonly, the cluster or manifold assumptions are used which imply that the regression function is smooth over high-density clusters or manifolds supporting the data. A generalization of these assumptions is that the regression function is smooth with respect to some density sensitive distance. This motivates the use of a density based metric [Bousquet et al., 2004, Coifman and Lafon, 2006, Sajama and Orlitsky, 2005] for semisupervised learning. We analyze this setting and make the following contributions - (a) we propose a semi-supervised learner that uses a density-sensitive kernel and show that it provides better performance than any supervised learner if the density support set has a small condition number and (b) we show that it is possible to adapt to the degree of semi-supervisedness using data-dependent choice of a parameter that controls sensitivity of the distance metric to the density. This ensures that the semisupervised learner never performs worse than a supervised learner even if the assumptions fail to hold.

## 1 Introduction

Semisupervised methods inevitably invoke some assumption that links the marginal distribution  $p(x)$  of the features  $X$  to the regression function  $f(x) = \mathbb{E}[Y|X = x]$  of the label  $Y$ . The most common assumption is the *cluster assumption* in which it is assumed that  $f$  is very smooth wherever  $p$  exhibits clusters [Lafferty and Wasserman, 2007, Rigollet, 2007, Seeger, 2000, Singh

et al., 2008a]. In the special case where the clusters are manifolds, this is called the *manifold assumption* [Belkin and Niyogi, 2004, Lafferty and Wasserman, 2007, Niyogi, 2008].

A generalization of the cluster and manifold assumptions is that the regression function is smooth with respect to some density-sensitive distance. Several recent papers propose using a density based metric or diffusion distance for semisupervised learning [Bousquet et al., 2004, Coifman and Lafon, 2006, Sajama and Orlitsky, 2005]. In this paper, we analyze semisupervised inference under this generalized assumption.

Singh, Nowak and Zhu [2008a], Lafferty and Wasserman [2007] and Nadler et al [2009] have showed that the degree to which unlabeled data improves performance is very sensitive to the cluster and manifold assumptions. In this paper, we introduce *adaptive semisupervised inference*. We define a parameter  $\alpha$  that controls the sensitivity of the distance metric to the density, and hence the strength of the semisupervised assumption. When  $\alpha = 0$  there is no semisupervised assumption, that is, there is no link between  $f$  and  $p$ . When  $\alpha = \infty$  there is a very strong semisupervised assumption. We use the data to estimate  $\alpha$  and hence we adapt to the appropriate assumption linking  $f$  and  $p$ .

This paper makes the following contributions - (a) we propose a semi-supervised learner that uses a density-sensitive kernel and show that it provides better performance than any supervised learner if the density support set has a small condition number and (b) we show that it is possible to adapt to the degree of semi-supervisedness using data-dependent choice of a parameter that controls sensitivity of the distance metric to the density. This ensures that the semisupervised learner never performs worse than a supervised learner even if the assumptions fail to hold. Preliminary simulations, to be reported in future work, confirmed that our proposed estimator adapts well to  $\alpha$  and has good risk when the semisupervised

smoothness holds and when it fails.

*Related Work.* There are a number of papers that discuss conditions under which semisupervised methods can succeed or that discuss metrics that are useful for semisupervised methods. These include Bousquet et al. [2004], Singh et al. [2008b], Nadler et al. [2009], Sajama and Orlitsky [2005] and references therein. However, to the best of our knowledge, there are no papers that explicitly study adaptive methods that allow the data to choose the strength of the semisupervised assumption.

*Outline.* This paper is organized as follows. In Section 2 we define a set of joint distributions  $\mathcal{P}_{XY}(\alpha)$  indexed by  $\alpha$ . In Section 3, we define a density sensitive estimator  $\hat{f}_\alpha$  of  $f$ , assuming that  $(f, p) \in \mathcal{P}_{XY}(\alpha)$ . We find finite sample bounds on the error of  $\hat{f}_\alpha$  and we investigate the dependence of this error on  $\alpha$ . In Section 4, we show that cross-validation can be used to adapt to  $\alpha$ . We conclude in Section 5.

## 2 Definitions

We consider the collection of joint distributions  $\mathcal{P}_{XY}(\alpha) = \mathcal{P}_X \times \mathcal{P}_{Y|X}$  indexed by a density-sensitivity parameter  $\alpha$  as follows.  $X, Y$  are random variables,  $X$  is supported on a compact domain  $\mathcal{X} \subset \mathbb{R}^d$ , and  $Y$  is real-valued. The marginal density  $p(x) \in [\lambda_0, \Lambda_0]$  is bounded over its support  $\{x : p(x) > 0\}$ , where  $0 < \lambda_0, \Lambda_0 < \infty$ . Also, let the conditional density be  $p(y|x)$  with variance bounded by  $\sigma^2$ , and conditional label mean or regression function be  $f(x) = \mathbb{E}[Y|X = x]$ , with  $|f(x)| \leq M$ . We say that  $(p, f) \in \mathcal{P}_{XY}(\alpha)$  if these functions satisfy the properties described below.

Before stating the properties of  $f$  and  $p$ , we define a distance metric with density sensitivity  $\alpha$ .

**Density-sensitive distance:** We consider the following distance with density sensitivity  $\alpha \in [0, \infty)$  between two points  $x_1, x_2 \in \mathcal{X}$  that is a modification of the definition in Sajama and Orlitsky [2005]:

$$D_\alpha(x_1, x_2) = \inf_{\gamma \in \Gamma(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt, \quad (1)$$

where  $\Gamma(x_1, x_2)$  is the set of all continuous finite curves from  $x_1$  to  $x_2$  with unit speed everywhere and  $L(\gamma)$  is the length of curve  $\gamma$  (i.e.  $\gamma(L(\gamma)) = x_2$ ). Notice that large  $\alpha$  makes points connected by high density paths closer, and  $\alpha = 0$  corresponds to Euclidean distance.

Our first assumption is that the regression function  $f$  is smooth with respect to the density sensitive distance:

**A1) Semisupervised smoothness:** The regression func-

tion  $f(x) = \mathbb{E}[Y|X = x]$  is  $\beta$ -smooth with respect to the density-sensitive distance  $D_\alpha$ , i.e. there exists constants  $C_1, \beta > 0$  such that for all  $x_1, x_2 \in \mathcal{X}$

$$|f(x_1) - f(x_2)| \leq C_1 [D_\alpha(x_1, x_2)]^\beta.$$

In particular if  $\alpha = 0$  and  $\beta = 1$ , this corresponds to Lipschitz smoothness.

Our second assumption is that the density function  $p$  is smooth with respect to Euclidean distance over the support set. Recall that the *support* of  $p$  is  $S = \{x : p(x) > 0\}$ .

**A2) Density smoothness:** The density function  $p(x)$  is Hölder  $\eta$ -smooth with respect to Euclidean distance if it has  $\lfloor \eta \rfloor$  derivatives and there exists a constant  $C_2 > 0$  such that for all  $x_1, x_2 \in S$

$$|p(x_1) - T_{x_2}^{[\eta]}(x_1)| \leq C_2 \|x_1 - x_2\|^\eta,$$

where  $\lfloor \eta \rfloor$  is the largest integer such that  $\lfloor \eta \rfloor < \eta$ , and  $T_{x_2}^{[\eta]}$  is the Taylor polynomial of degree  $\lfloor \eta \rfloor$  around the point  $x_2$ .

The *condition number* of a set  $S$  with boundary  $\partial S$  is the largest real number  $\tau > 0$  such that, if  $d(x, \partial S) \leq \tau$  then  $x$  has a unique projection onto the boundary of  $S$ . Here,  $d(x, \partial S) = \inf_{z \in \partial S} \|x - z\|$ . When  $\tau$  is large,  $S$  cannot be too thin, the boundaries of  $S$  cannot be too curved and  $S$  cannot get too close to being self-intersecting. If  $S$  consists of more than one connected component, then  $\tau$  large also means that the connected components cannot be too close to each other. Let  $\tau_0$  denote the smallest condition number of the support sets  $S$  of all  $p \in \mathcal{P}_X$ . We shall see that semisupervised inference outperforms supervised inference when  $\tau_0$  is small. Additionally, we assume that  $S$  has at most  $K < \infty$  connected components.

In the supervised setting, we assume access to  $n$  labeled data  $\mathcal{L} = \{X_i, Y_i\}_{i=1}^n$  drawn i.i.d. from  $\mathcal{P}_{XY}(\alpha)$ , and in the semi-supervised setting, we assume access to  $m$  additional unlabeled data  $\mathcal{U} = \{X_i\}_{i=1}^m$  drawn i.i.d. from  $\mathcal{P}_X$ .

As usual, we write  $a_n = O(b_n)$  if  $|a_n/b_n|$  is bounded for all large  $n$ . Similarly,  $a_n = \Omega(b_n)$  if  $|a_n/b_n|$  is bounded away from 0 for all large  $n$ . We write  $a_n \asymp b_n$  if  $a_n = O(b_n)$  and  $a_n = \Omega(b_n)$ .

## 3 Density-Sensitive Inference

Let  $K(x)$  be a symmetric non-negative function and let  $K_h(x) = K(\|x\|/h)$ . Let

$$\hat{p}_m(x) = \frac{1}{m} \sum_{i=1}^m \frac{1}{h_m^d} K_{h_m}(x - X_i) \quad (2)$$

be the kernel density estimator of  $p$  with bandwidth  $h_m$ , based on the unlabeled data. Define the support set estimate  $\hat{S} = \{x : \hat{p}_m(x) > 0\}$  and the empirical boundary region

$$\hat{\mathcal{R}}_{\hat{S}} = \left\{ x : \inf_{z \in \partial \hat{S}} \|x - z\|_2 < 2\delta_m \right\}.$$

where  $\delta_m = 2c_2\sqrt{d}((\log^2 m)/m)^{\frac{1}{d}}$  for some constant  $c_2 > 0$ . Now define a plug-in estimate of the  $D_\alpha$  distance as follows:

$$\hat{D}_{\alpha,m}(x_1, x_2) = \inf_{\gamma \in \hat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{\hat{p}_m(\gamma(t))^\alpha} dt,$$

where  $\hat{\Gamma}(x_1, x_2) = \{\gamma \in \Gamma(x_1, x_2) : \forall t \in [0, L(\gamma)] \gamma(t) \in \hat{S} \setminus \hat{\mathcal{R}}_{\hat{S}}\}$ , and  $\hat{D}_{\alpha,m}(x_1, x_2) = \infty$  if  $\hat{\Gamma}(x_1, x_2) = \emptyset$ .

We consider the following semisupervised learner which uses a kernel that is sensitive to the density. In the following definitions we take, for simplicity,  $K(x) = I(\|x\| \leq 1)$ .

**Semisupervised kernel estimator:**

$$\hat{f}_{h,\alpha}(x) = \frac{\sum_{i=1}^n Y_i K_h(\hat{D}_{\alpha,m}(x, X_i))}{\sum_{i=1}^n K_h(\hat{D}_{\alpha,m}(x, X_i))}. \quad (3)$$

### 3.1 Performance upper bound for semisupervised estimator

The following theorem characterizes the performance of the density sensitive semisupervised kernel estimator.

**Theorem 1.** Assume  $\lambda_0 > 1 + c_0$  for some constant  $c_0 > 0$ <sup>1</sup> and let  $\epsilon_m = c_1(\log m)^{-1/2}$  for constant  $c_1 > 0$  and  $\delta_m = 2c_2\sqrt{d}((\log^2 m)/m)^{\frac{1}{d}}$  for some constant  $c_2 > 0$ . If  $\tau_0 \in (3\delta_m, \infty)$  and  $h > (2c_4/(\tau_0^{d-1}(\lambda_0 - \epsilon_m)^\alpha))$  where  $c_4 > 0$  is a constant, then for large enough  $m$

$$\begin{aligned} \sup_{(p,f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_{n,m} \left\{ \int (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} &\leq \\ &(M^2 + \sigma^2) \left( \frac{1}{m} + 3c_3 2^d \Lambda_0 \frac{\delta_m}{\tau_0} \right) \\ &+ \left[ h \left( \frac{\lambda_0 + \epsilon_m}{\lambda_0} \right)^\alpha \right]^{2\beta} \\ &+ \frac{K(M^2/e + 2\sigma^2)}{n}. \end{aligned}$$

<sup>1</sup>This assumption is more restrictive than necessary, and a more general statement can be by introducing a rescaling factor in the definition of the density-sensitive distance.

The proof of Theorem 1 is given in section 6. The first term is negligible when the amount of unlabeled data  $m$  is large. The second term is the bias and third term is variance. If the bandwidth

$$h \asymp \frac{1}{\delta_m^{d-1} \lambda_0^\alpha}$$

and  $\alpha \asymp \log m$  is large enough, then the density-sensitive semisupervised kernel estimator is able to achieve an integrated MSE rate of  $O(n^{-1})$  for all joint distributions in  $\mathcal{P}_{XY}(\alpha)$  supported on sets with condition number  $\tau_0 > 3\delta_m$ .

### 3.2 Performance lower bound for any supervised estimator

We now establish a lower bound on the performance of any supervised estimator.

**Theorem 2.** Assume  $d \geq 2$  and  $\alpha > 0$ . There exists a constant  $c_5 > 0$  depending only on  $d$  so that if  $\tau_0 \leq c_5 n^{-\frac{1}{d-1}}$ , then

$$\inf_{\hat{f}} \sup_{(p,f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_n \int (\hat{f}(x) - f(x))^2 dP(x) = \Omega(1)$$

where the inf is over all supervised estimators.

Coupled with Theorem 1, the results state that if the condition number of the support set is small  $3\delta_m < \tau_0 \leq c_5 n^{-\frac{1}{d-1}}$  and  $\alpha$  is large enough, then the density-sensitive semi-supervised estimator outperforms any supervised learning algorithm in terms of integrated MSE rate.

A complete proof of Theorem 2 is given in the appendix. Here we provide some intuition regarding the proof strategy. We construct a set of joint distributions over  $X$  and  $Y$  that depends on  $n$ , and apply Assouad's Lemma. Intuitively, we need to take advantage of the decreasing condition number  $\tau_0$ . This is because if  $\tau_0$  were to be kept fixed, as  $n$  increases the semi-supervised assumption would reduce to familiar Euclidean smoothness.

So, we construct the distributions as follows. We split the unit cube in  $\mathbb{R}^d$  into two rectangle sets with a small gap in between, and let the marginal density  $p$  be uniform over these sets. Then we add a series of ‘‘bumps’’ between the two rectangles, as shown schematically in Figure 1. Over one of the sets we set  $f \equiv M$ , and over the other we set  $f \equiv -M$ . The number of bumps increases with  $n$ , implying that the condition number must decrease. The sets are designed specifically so that the condition number can be lower bounded easily as a function of  $n$ . In essence, as  $n$  increases these boundaries become space-filling, so that there is a region where the regression function could be  $M$  or  $-M$ , and it is not possible to tell which with only labeled data.

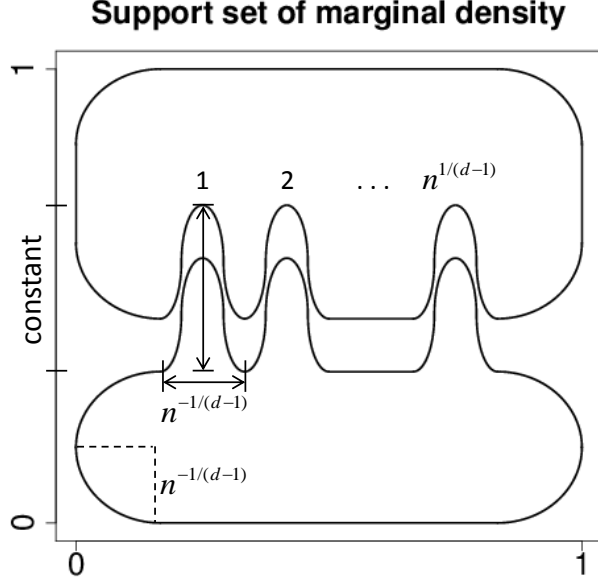


Figure 1: A two-dimensional cross-section of the support of a marginal density  $p$  used in the proof of Theorem 2.

## 4 Adaptive Semisupervised Inference

In section 3.1, we established a bound on the integrated mean square error of the density-sensitive semisupervised kernel estimator. The bound is achieved by using an estimate  $\hat{D}_\alpha$  of the density-sensitive distance. However, this requires knowing the density-sensitive parameter  $\alpha$ , along with other parameters.

It is critical to choose  $\alpha$  (and  $h$ ) appropriately, otherwise we might incur a large error if the semisupervised assumption does not hold or holds with a different density sensitivity value  $\alpha$ . The following result shows that we can adapt to the correct degree of semisupervisedness  $\alpha$  if cross-validation is used to select the appropriate  $\alpha$  and  $h$ . This implies that the estimator gracefully degrades to a supervised learner if the semisupervised assumption (sensitivity of regression function to marginal density) does not hold ( $\alpha = 0$ ).

For any  $f$ , define the risk  $R(f) = \mathbb{E}[(f(X) - Y)^2]$  and the excess risk  $\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$  where  $f^*$  is the true regression function. Let  $\mathcal{H}$  be a finite set of bandwidths and let  $\mathcal{A}$  be a finite set of values for  $\alpha$ . Divide the data into training data  $T$  and validation data  $V$ . For notational simplicity, let both sets have size  $n$ . Let  $\mathcal{F} = \{\hat{f}_{\alpha,h}^T\}_{\alpha \in \mathcal{A}, h \in \mathcal{H}}$  denote the semisupervised kernel estimators trained on data  $T$  using  $\alpha \in \mathcal{A}$  and  $h \in \mathcal{H}$ . For each  $\hat{f}_{\alpha,h}^T \in \mathcal{F}$  let  $\hat{R}^V(\hat{f}_{\alpha,h}^T) = n^{-1} \sum_{i=1}^n (\hat{f}_{\alpha,h}^T(X_i) - Y_i)^2$  where the sum is over  $V$ . Let  $Y_i = f(X_i) + \epsilon_i$  with  $\epsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ . Also, we assume that  $|f(x)|, |\hat{f}_{\alpha,h}^T(x)| \leq M$ , where

$M > 0$  is a constant.<sup>2</sup>

**Theorem 3.** Let  $\mathcal{F} = \{\hat{f}_{\alpha,h}^T\}_{\alpha \in \mathcal{A}, h \in \mathcal{H}}$  denote the semisupervised kernel estimators trained on data  $T$  using  $\alpha \in \mathcal{A}$  and  $h \in \mathcal{H}$ . Use validation data  $V$  to pick

$$(\hat{\alpha}, \hat{h}) = \arg \min_{(\alpha \in \mathcal{A}, h \in \mathcal{H})} \hat{R}^V(\hat{f}_{\alpha,h}^T)$$

and define the corresponding estimator  $\hat{f}_{\hat{\alpha}, \hat{h}}$ . Then, for every  $0 < \delta < 1$ ,

$$\mathbb{E}[\mathcal{E}(\hat{f}_{\hat{\alpha}, \hat{h}})] \leq \frac{1}{1-a} \left[ \min_{\alpha \in \mathcal{A}, h \in \mathcal{H}} \mathbb{E}[\mathcal{E}(\hat{f}_{\alpha,h})] + \frac{\log(|\mathcal{A}||\mathcal{H}|/\delta)}{nt} \right] + 4\delta M^2$$

where  $0 < a < 1$  and  $0 < t < 15/(38(M^2 + \sigma^2))$  are constants.  $\mathbb{E}$  denotes expectation over everything that is random.

See appendix for proof. In practice, both  $\mathcal{H}$  and  $\mathcal{A}$  may be taken to be of size  $n^a$  for some  $a > 0$ . Then we can approximate the optimal  $h$  and  $\alpha$  with sufficient accuracy to achieve the optimal rate. Setting  $\delta = 1/(4M^2n)$ , we then see that the penalty for adaptation is  $\frac{\log(|\mathcal{A}||\mathcal{H}|/\delta)}{nt} + \delta M = O(\log n/n)$  and hence introduces only a logarithmic term.

## 5 Discussion

Semisupervised methods are very powerful but, like all methods, they only work under certain conditions.

We have shown that, when the support of the distribution is somewhat irregular (i.e., the boundary of the support of the density has a small condition number), then semi-supervised methods can attain better performance. Specifically, we demonstrated that a semi-supervised kernel estimator that uses a density-sensitive distance can outperform any supervised estimator in such cases.

We introduced a family of estimators indexed by a parameter  $\alpha$ . This parameter controls the strength of the semi-supervised assumption. We showed that the behavior of the semi-supervised method depends critically on  $\alpha$ .

Finally, we showed that cross-validation can be used to automatically adapt to  $\alpha$  so that  $\alpha$  does not need to be known. Hence, our method takes advantage of the unlabeled data when the semi-supervised assumption holds, but does not add extra bias when the assumption fails. Preliminary simulations confirm that our proposed estimator adapts well to  $\alpha$  and has good risk when the

<sup>2</sup> Note that the estimator can always be truncated if necessary.

semi-supervised smoothness holds and when it fails. We will report these results in future work.

The analysis in this paper can be extended in several ways. First, it is possible to use other density sensitive metrics such as the diffusion distance [Lee and Wasserman, 2008]. Second, it is possible to relax the assumption that the density  $p$  is strictly bounded away from 0 on its support. Finally, other estimators besides kernel estimators can be used. We will report on these extensions elsewhere.

## 6 Proof of Theorem 1

Here we prove Theorem 1 stated in section 3.1 (repeated below for convenience), using some results given in the appendix.

**Theorem 4.** Assume  $\lambda_0 > 1 + c_0$  for some constant  $c_0 > 0$ <sup>3</sup> and let  $\epsilon_m = c_1(\log m)^{-1/2}$  for constant  $c_1 > 0$  and  $\delta_m = 2c_2\sqrt{d}((\log^2 m)/m)^{\frac{1}{d}}$  for some constant  $c_2 > 0$ . If  $\tau_0 \in (3\delta_m, \infty)$  and  $h > (2c_4/(\tau_0^{d-1}(\lambda_0 - \epsilon_m)^\alpha))$  where  $c_4 > 0$  is a constant, then for large enough  $m$

$$\begin{aligned} \sup_{(p,f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_{n,m} \left\{ \int (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} &\leq \\ &(M^2 + \sigma^2) \left( \frac{1}{m} + 3c_3 2^d \Lambda_0 \frac{\delta_m}{\tau_0} \right) \\ &+ \left[ h \left( \frac{\lambda_0 + \epsilon_m}{\lambda_0} \right)^\alpha \right]^{2\beta} \\ &+ \frac{K(M^2/e + 2\sigma^2)}{n}. \end{aligned}$$

*Proof.* Let  $\mathcal{G}_m$  be the indicator of the event when the unlabeled sample is such that  $\sup_{x \in S \setminus \mathcal{R}_{\partial S}} |p(x) - \hat{p}_m(x)| \leq \epsilon_m$  and  $\partial \hat{S} \subset \mathcal{R}_{\partial S}$ . From Theorem 5,

$$\begin{aligned} \mathbb{E}_{n,m} \left\{ (1 - \mathcal{G}_m) \int (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ \leq \frac{1}{m} (M^2 + \sigma^2). \end{aligned}$$

We can write

$$\begin{aligned} \mathbb{E}_{n,m} \left\{ \mathcal{G}_m \int (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ = \mathbb{E}_{n,m} \left\{ \mathcal{G}_m \int_{S_m^*} (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ + \mathbb{E}_{n,m} \left\{ \mathcal{G}_m \int_{S \setminus S_m^*} (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \end{aligned}$$

where  $S_m^*$  as defined in Proposition 2. For the boundary region we have

$$\begin{aligned} \mathbb{E}_{n,m} \left\{ \mathcal{G}_m \int_{S \setminus S_m^*} (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ \leq (M^2 + \sigma^2) P(S \setminus S_m^*) \\ \leq \Lambda_0 (M^2 + \sigma^2) \text{Leb}(S \setminus S_m^*) \end{aligned}$$

where  $\text{Leb}$  denotes the Lebesgue measure. Since the radius of curvature of  $\partial S$  is at least  $\tau_0$ , and  $\tau_0 > 3\delta_m$ , we have by Proposition 3,

$$\begin{aligned} \text{Leb}(S \setminus S_m^*) &\leq \text{Vol}(\partial S) \frac{(\tau_0 + 3\delta_m)^d - \tau_0^d}{\tau_0^{d-1}} \\ &\leq c_3 \left[ \left( 1 + \frac{3\delta_m}{\tau_0} \right)^d - 1 \right] \\ &\leq c_3 \sum_{i=1}^d \binom{d}{i} \frac{3\delta_m}{\tau_0} \\ &\leq 3c_3 2^d \frac{\delta_m}{\tau_0} \end{aligned}$$

where  $\text{Vol}$  denotes the  $d-1$ -dimensional volume on  $\partial S$ . So

$$\begin{aligned} \mathbb{E}_{n,m} \left\{ \mathcal{G}_m \int_{S \setminus S_m^*} (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ \leq 3c_3 2^d \Lambda_0 (M^2 + \sigma^2) \frac{\delta_m}{\tau_0}. \end{aligned}$$

Following the derivation in Chapter 5 of Györfi et al. [2002], we have

$$\begin{aligned} \mathbb{E}_n \left\{ \mathcal{G}_m \int_{S_m^*} (\hat{f}_{h,\alpha}(x) - f(x))^2 dP(x) \right\} \\ \leq \mathcal{G}_m C_1^2 \sup_{x \in S_m^*} \sup_{x' \in S \cap S_{x,h}^{\hat{D}_{\alpha,m}}} D_\alpha(x, x')^{2\beta} \\ + \mathcal{G}_m \frac{M^2/e + 2\sigma^2}{n} \mathcal{N} \left( S_m^*, \hat{D}_{\alpha,m}, \frac{h}{2} \right) \end{aligned}$$

where  $S_{x,h}^{\hat{D}_{\alpha,m}} = \{x' : \hat{D}_{\alpha,m}(x, x') \leq h\}$ , and  $\mathcal{N}$  denotes the covering number. Note that since  $\hat{\Gamma}(x, x') = \emptyset \Rightarrow \hat{D}_{\alpha,m} = \infty$ , we will always have  $(x, x') \in \Psi$  if  $x' \in S \cap S_{x,h}^{\hat{D}_{\alpha,m}}$  (and, of course, the same applies when  $x' \in S_m^* \cap S_{x,h/2}^{\hat{D}_{\alpha,m}}$ ). So we can apply Proposition 2 to give

$$\mathcal{G}_m \sup_{x \in S_m^*} \sup_{x' \in S \cap S_{x,h}^{\hat{D}_{\alpha,m}}} D_\alpha(x, x')^{2\beta} \leq \left[ h \left( \frac{\lambda_0 + \epsilon_m}{\lambda_0} \right)^\alpha \right]^{2\beta}$$

<sup>3</sup>This assumption is more restrictive than necessary, and a more general statement can be by introducing a rescaling factor in the definition of the density-sensitive distance.

and

$$\mathcal{G}_m \mathcal{N} \left( S_m^*, \hat{D}_{\alpha, m}, \frac{h}{2} \right) \leq \mathcal{G}_m \mathcal{N} \left( S_m^*, d_{S_m^*}, \frac{h(\lambda_0 - \epsilon_m)^\alpha}{2} \right)$$

where the  $d_{S_m^*}$  distance is the length of the shortest path between two points restricted to  $S_m^*$ , as defined in the appendix. Clearly  $S_m^*$  has condition number at least  $\tau_0 - 3\delta_m > 0$ . If  $S_m^*$  has exactly one connected component, then Proposition 4 combined with the assumption that  $h > (2c_4/(\tau_0^{d-1}(\lambda_0 - \epsilon_m)^\alpha))$  implies that any point in  $S_m^*$  is a  $h(\lambda_0 - \epsilon_m)^\alpha/2$  covering, so

$$\mathcal{N} \left( S_m^*, d_{S_m^*}, \frac{h(\lambda_0 - \epsilon_m)^\alpha}{2} \right) = 1.$$

Since  $S_m^*$  can have at most  $K$  connected components, we can repeat the same argument for each component and conclude that

$$\mathcal{N} \left( S_m^*, d_{S_m^*}, \frac{h(\lambda_0 - \epsilon_m)^\alpha}{2} \right) \leq K.$$

So,

$$\begin{aligned} & \mathbb{E}_{n, m} \left\{ \int (\hat{f}_{h, \alpha}(x) - f(x))^2 dP(x) \right\} \\ & \leq (M^2 + \sigma^2) \left( \frac{1}{m} + 3c_3 2^d \Lambda_0 \frac{\delta_m}{\tau_0} \right) \\ & \quad + \left[ h \left( \frac{\lambda_0 + \epsilon_m}{\lambda_0} \right)^\alpha \right]^{2\beta} \\ & \quad + \frac{K(M^2/e + 2\sigma^2)}{n}. \end{aligned}$$

□

## Acknowledgments

This research is supported in part by AFOSR under grants FA9550-10-1-0382 and FA95500910373 and NSF under grants IIS-1116458 and DMS-0806009.

## References

- M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56(1-3): 209–239, 2004.
- O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In *Advances in Neural Information Processing Systems*, 2004.
- R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.
- C. Genovese, M. Perone-Pacífico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *Arxiv preprint arXiv:1007.0549*, 2010.
- L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Verlag, 2002.
- J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Trans. Info. Th.*, 52(9): 4036–4048, 2006.
- J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems 20*, pages 801–808, 2007.
- A. B. Lee and L. Wasserman. Spectral Connectivity Analysis. *Arxiv preprint arXiv:0811.0121*, 2008.
- B. Nadler, N. Srebro, and X. Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems 22*, pages 1330–1338, 2009.
- P. Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical Report TR-2008-01, Computer Science Department, University of Chicago. URL <http://people.cs.uchicago.edu/~niyogi/papersps/ssminimax2.pdf>, 2008.
- P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.
- P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.
- Sajama and A. Orlitsky. Estimating and computing density based distance metrics. In *Proceedings of the 22nd international conference on Machine learning, ICML 2005*, pages 760–767, 2005.
- M. Seeger. Learning with labeled and unlabeled data. Technical report, Institute for ANC, Edinburgh, UK. URL <http://citeseer.ist.psu.edu/seeger01learning.html>, 2000.
- A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn’t. In *Neural Information Processing Systems (NIPS)*, 2008a.
- A. Singh, R. D. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn’t. Technical report, University of Wisconsin - Madison, ECE Department. URL [http://www.cae.wisc.edu/~singh/SSL\\_TR.pdf](http://www.cae.wisc.edu/~singh/SSL_TR.pdf), 2008b.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

## Appendix

**Results used in proof of Theorem 1** In order to prove Theorem 1, we characterize how the plug-in density-sensitive distance estimate  $\widehat{D}_\alpha$  behaves. For this, we start with a result about the density estimator.

**Theorem 5.** *If  $m \geq m_0$ , where  $m_0 \equiv m_0(\lambda_0, \Lambda_0)$  is a constant, then for all marginal densities  $p$  of distributions in  $\mathcal{P}_{XY}(\alpha)$ , we have with probability  $> 1 - 1/m$ ,*

$$\sup_{x \in S \setminus \mathcal{R}_{\partial S}} |p(x) - \widehat{p}_m(x)| \leq \epsilon_m \text{ and } \partial \widehat{S} \subset \mathcal{R}_{\partial S}$$

where  $\epsilon_m = c_1(\log m)^{-1/2}$  for constant  $c_1 \equiv c_1(K, C_2, d, \eta, \Lambda_0)$ ,  $\widehat{S} = \{x : \widehat{p}_m(x) > 0\}$ , and

$$\mathcal{R}_{\partial S} = \left\{ x : \inf_{z \in \partial S} \|x - z\|_2 < \delta_m \right\}$$

where  $\delta_m = 2c_2\sqrt{d} \left( \frac{\log^2 m}{m} \right)^{\frac{1}{d}}$  for some constant  $c_2 > 0$ .

*Proof.* Follows from Theorem 1 in Singh et al. [2008a] by noting that since the density estimate will be 0 a.s. outside the boundary region, and we have  $p \geq \lambda_0$  on  $S$ , for sufficiently large  $m$  (i.e. small  $\epsilon_m$ ), we must have  $S \setminus \mathcal{R}_{\partial S} \subseteq \widehat{S} \subseteq S \cup \mathcal{R}_{\partial S}$ .  $\square$

The following two propositions now characterize how the plug-in density-sensitive distance estimate  $\widehat{D}_\alpha$  behaves.

**Proposition 1.** *Assume  $\sup_{x \in S \setminus \mathcal{R}_{\partial S}} |\widehat{p}_m(x) - p(x)| \leq \epsilon_m$  and  $\partial \widehat{S} \subset \mathcal{R}_{\partial S}$ . Let*

$$\widetilde{D}_{\alpha,m}(x_1, x_2) = \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt$$

and  $\Psi = \{(x_1, x_2) : x_1, x_2 \in \widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S}, \widehat{\Gamma}(x_1, x_2) \neq \emptyset\}$ . Then for any  $(x_1, x_2) \in \Psi$ ,

$$\begin{aligned} \left( \frac{\lambda_0}{\lambda_0 + \epsilon_m} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2) &\leq \widehat{D}_{\alpha,m}(x_1, x_2) \\ &\leq \left( \frac{\lambda_0}{(\lambda_0 - \epsilon_m)_+} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2). \end{aligned}$$

*Proof.* Note that by the triangle inequality,  $\mathcal{R}_{\partial S} \subseteq \widehat{\mathcal{R}}_{\partial S}$ , so  $\widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S} \subseteq S \setminus \mathcal{R}_{\partial S}$  since  $\tau_0 > 2\delta_m$  for  $m$  large enough. We see that if  $(x_1, x_2) \in \Psi$ , then  $x$  and  $y$  must be in the same connected component of  $\widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S}$ , and, furthermore, all points along any path in  $\widehat{\Gamma}(x_1, x_2)$

must also be in the same connected component. For  $(x_1, x_2) \in \Psi$ ,

$$\begin{aligned} \widehat{D}_{\alpha,m}(x_1, x_2) &= \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} \frac{p(\gamma(t))^\alpha}{\widehat{p}_m(\gamma(t))^\alpha} dt \\ &\leq \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \left[ \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt \right] \left[ \sup_{t \in [0, L(\gamma)]} \left( \frac{p(\gamma(t))}{\widehat{p}_m(\gamma(t))} \right)^\alpha \right] \\ &\leq \sup_{z \in S \setminus \mathcal{R}_{\partial S}} \left( \frac{p(z)}{\widehat{p}_m(z)} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2) \end{aligned}$$

and

$$\begin{aligned} \sup_{z \in S \setminus \mathcal{R}_{\partial S}} \left( \frac{p(z)}{\widehat{p}_m(z)} \right)^\alpha &\leq \sup_{z \in S \setminus \mathcal{R}_{\partial S}} \left( \frac{p(z)}{(p(z) - \epsilon_m)_+} \right)^\alpha \\ &\leq \left( \frac{\lambda_0}{(\lambda_0 - \epsilon_m)_+} \right)^\alpha. \end{aligned}$$

So

$$\widehat{D}_{\alpha,m}(x_1, x_2) \leq \left( \frac{\lambda_0}{(\lambda_0 - \epsilon_m)_+} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2).$$

Similarly,

$$\begin{aligned} \widehat{D}_{\alpha,m}(x_1, x_2) &\geq \inf_{z \in S \setminus \mathcal{R}_{\partial S}} \left( \frac{p(z)}{p(z) + \epsilon_m} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2) \\ &\geq \left( \frac{\lambda_0}{\lambda_0 + \epsilon_m} \right)^\alpha \widetilde{D}_{\alpha,m}(x_1, x_2). \end{aligned}$$

$\square$

Given a set  $A \subseteq \mathbb{R}^d$ , define

$$d_A(x_1, x_2) = \inf_{\gamma \in \Gamma_A(x_1, x_2)} L(\gamma)$$

where  $\Gamma_A(x_1, x_2) = \{\gamma \in \Gamma(x_1, x_2) : \forall t \in [0, L(\gamma)] \gamma(t) \in A\}$ .

**Proposition 2.** *With the notation of Proposition 1, for all  $x_1, x_2$ ,*

$$D_{\alpha,m}(x_1, x_2) \leq \widetilde{D}_{\alpha,m}(x_1, x_2).$$

Assume  $\sup_{x \in S \setminus \mathcal{R}_{\partial S}} |\widehat{p}_m(x) - p(x)| \leq \epsilon_m$  and  $\partial \widehat{S} \subset \mathcal{R}_{\partial S}$ . Then for any  $(x_1, x_2) \in \Psi$ ,

$$\widetilde{D}_{\alpha,m}(x_1, x_2) \leq \frac{d_{S \setminus \widehat{\mathcal{R}}_{\partial S}}(x_1, x_2)}{\lambda^\alpha}$$

and

$$\left(\frac{\lambda}{\lambda + \epsilon_m}\right)^\alpha D_\alpha(x_1, x_2) \leq \widehat{D}_{\alpha, m}(x_1, x_2) \leq \frac{d_{S_m^*}(x_1, x_2)}{(\lambda_0 - \epsilon_m)_+^\alpha}$$

$$\text{where } S_m^* = \left\{x \in S : \inf_{z \in \partial S} \|x - z\|_2 \geq 3\delta_m\right\}.$$

*Proof.* Since for any  $x_1$  and  $x_2$ ,  $\widehat{\Gamma}(x_1, x_2) \subseteq \Gamma(x_1, x_2)$ , clearly  $D_{\alpha, m}(x_1, x_2) \leq \widehat{D}_{\alpha, m}(x_1, x_2)$ . If  $\partial \widehat{S} \subset \mathcal{R}_{\partial S}$ , write

$$\begin{aligned} \widehat{D}_{\alpha, m}(x_1, x_2) &= \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} \frac{1}{p(\gamma(t))^\alpha} dt \\ &\leq \left[ \sup_{z \in \widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S}} \frac{1}{p(z)^\alpha} \right] \left[ \inf_{\gamma \in \widehat{\Gamma}(x_1, x_2)} \int_0^{L(\gamma)} dt \right] \\ &\leq \frac{1}{\lambda_0^\alpha} d_{\widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S}}(x_1, x_2) \\ &\leq \frac{1}{\lambda_0^\alpha} d_{S_m^*}(x_1, x_2) \end{aligned}$$

since, by the triangle inequality,  $S_m^* \subseteq \widehat{S} \setminus \widehat{\mathcal{R}}_{\partial S}$ . Applying Proposition 1, the result follows.  $\square$

To prove Theorem 1, we also need the following two results.

**Proposition 3.** *Let  $\mathcal{X}$  be a compact subset of  $\mathbb{R}^d$ , and  $T > 0$ . Then for any  $\tau \in (0, T)$ , for all sets  $S \subseteq \mathcal{X}$  with condition number at least  $\tau$ ,  $\text{Vol}(\partial S) \leq c_3/\tau$  for some  $c_3$  independent of  $\tau$ , where  $\text{Vol}$  is the  $d - 1$ -dimensional volume.*

*Proof.* Let  $\{z_i\}_{i=1}^N$  be a minimal Euclidean  $\tau/2$ -covering of  $\partial S$ , and  $B_i = \{x : \|x - z_i\|_2 \leq \tau/2\}$ . Let  $T_i$  be the tangent plane to  $\partial S$  at  $z_i$ . Then using the argument made in the proof of Lemma 4 in Genovese et al. [2010],

$$\begin{aligned} \text{Vol}(B_i \cap \partial S) &\leq C_1 \text{Vol}(B_i \cap T_i) \frac{1}{\sqrt{1 - (\tau/2)^2/\tau^2}} \\ &\leq C_2 \tau^{d-1} \end{aligned}$$

for some constants  $C_1$  and  $C_2$  independent of  $\tau$ . Since  $\mathcal{X}$  is compact,

$$\mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \leq C \left(\frac{1}{\tau}\right)^d$$

for some constant  $C$  depending only on  $\mathcal{X}$  and  $T$ , where  $\mathcal{N}$  denotes the covering number (note that even though  $\partial S$  is a  $d - 1$  dimensional set, we can't claim  $\mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \leq C \left(\frac{1}{\tau}\right)^{d-1}$ ).

$\|\cdot\|_2, \tau) = O(\tau^{-(d-1)})$ , since  $\partial S$  can become space-filling as  $\tau \rightarrow 0$ ). So

$$\begin{aligned} \text{Vol}(\partial S) &\leq \sum_{i=1}^N \text{Vol}(B_i \cap \partial S) \\ &\leq C_2 \tau^{d-1} \mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \\ &\leq C_2 C \tau^{-1} \end{aligned}$$

and the result follows with  $c_3 = C_2 C$ .  $\square$

**Proposition 4.** *Let  $\mathcal{X}$  be a compact subset of  $\mathbb{R}^d$ , and  $T > 0$ . Then for any  $\tau \in (0, T)$ , for all compact, connected sets  $S \subseteq \mathcal{X}$  with condition number at least  $\tau$ ,  $\sup_{u, v \in S} d_S(u, v) \leq c_4 \tau^{1-d}$  for some  $c_4$  independent of  $\tau$ .*

*Proof.* First consider the quantity

$$\sup_{u, v \in \partial S} d_S(u, v).$$

Since  $\partial S \subseteq S$ , clearly

$$\sup_{u, v \in \partial S} d_S(u, v) \leq \sup_{u, v \in \partial S} d_{\partial S}(u, v).$$

Since  $\partial S$  is closed, there must exist  $u^*, v^* \in \partial S$  such that

$$\sup_{u, v \in \partial S} d_{\partial S}(u, v) = d_{\partial S}(u^*, v^*).$$

Let  $\{z_i\}_{i=1}^N$  be a minimal  $\tau$ -covering of  $\partial S$  in the  $d_{\partial S}$  metric. Let  $\{\tilde{z}_i\}_{i=1}^{\tilde{N}} \subseteq \{z_i\}_{i=1}^N$  such that  $d_{\partial S}(u^*, \tilde{z}_1) \leq \tau$ ,  $d_{\partial S}(v^*, \tilde{z}_{\tilde{N}}) \leq \tau$ , and for any  $1 \leq i \leq \tilde{N} - 1$ ,  $d_{\partial S}(\tilde{z}_i, \tilde{z}_{i+1}) \leq 2\tau$ . Then

$$\begin{aligned} d_{\partial S}(u^*, v^*) &\leq d_{\partial S}(u^*, \tilde{z}_1) + d_{\partial S}(v^*, \tilde{z}_{\tilde{N}}) \\ &\quad + \sum_{i=1}^{\tilde{N}-1} d_{\partial S}(\tilde{z}_i, \tilde{z}_{i+1}) \\ &\leq 2\tau \tilde{N}. \end{aligned}$$

So,

$$d_{\partial S}(u^*, v^*) \leq 2\tau \mathcal{N}(\partial S, d_{\partial S}, \tau).$$

By Proposition 6.3 in Niyogi et al. [2008] (or see Lemma 3 in Genovese et al. [2010]), if  $x, y \in \partial S$  such that  $\|x - y\|_2 = a \leq \tau/2$ , then  $d_{\partial S}(x, y) \leq \tau - \tau\sqrt{1 - (2a)/\tau}$ . In particular, if  $\|x - y\|_2 \leq \tau/2$ , then  $d_{\partial S}(x, y) \leq \tau$ . So any Euclidean  $\tau/2$ -covering of  $\partial S$  is also a  $\tau$ -covering



in the  $d_{\partial S}$  metric. Then we have

$$\begin{aligned}
\sup_{u,v \in \partial S} d_S(u,v) &\leq d_{\partial S}(u^*, v^*) \\
&\leq 2\tau \mathcal{N}(\partial S, d_{\partial S}, \tau) \\
&\leq 2\tau \mathcal{N}(\partial S, \|\cdot\|_2, \tau/2) \\
&\leq C\tau \left(\frac{1}{\tau}\right)^d \\
&= C\tau^{1-d}
\end{aligned}$$

for some constant  $C$  depending only on  $\mathcal{X}$  and  $T$  (note that, as in the proof of 3, even though  $\partial S$  is a  $d-1$  dimensional set, we can't claim  $\mathcal{N}(\partial S, \|\cdot\|_2, \tau) = O(\tau^{-(d-1)})$ , since  $\partial S$  can become space-filling as  $\tau \rightarrow 0$ ).

Now let  $u^\dagger, v^\dagger \in S$  such that

$$\sup_{u,v \in S} d_S(u,v) = d_S(u^\dagger, v^\dagger)$$

which must exist since  $S$  is compact. Let  $u^\ddagger, v^\ddagger \in \partial S$  be the (not necessarily unique) projections of  $u^\dagger$  and  $v^\dagger$  onto  $\partial S$ . Clearly the line segment connecting  $u^\dagger$  and  $u^\ddagger$  is fully contained in  $S$ , and the same applies to  $v^\dagger$  and  $v^\ddagger$ . So

$$\begin{aligned}
d_S(u^\dagger, v^\dagger) &\leq d_S(u^\dagger, u^\ddagger) + d_S(u^\ddagger, v^\ddagger) + d_S(v^\ddagger, v^\dagger) \\
&\leq \|u^\dagger - u^\ddagger\|_2 + \|v^\dagger - v^\ddagger\|_2 + d_S(u^*, v^*) \\
&\leq 2 \text{diam}(\mathcal{X}) + C\tau^{1-d}
\end{aligned}$$

and setting  $c_4 = 2T^{d-1} \text{diam}(\mathcal{X}) + C$ , the result follows.  $\square$

**Proof of Theorem 2** The proof of Theorem 2 is based on the following result based on Assouad's Lemma (see e.g. Tsybakov [2009]).

**Theorem 6.** Let  $\Omega = \{0,1\}^q$ , the collection of binary vectors of length  $q \geq 1$ . Let  $\mathcal{P}_\Omega = \{P^\omega, \omega \in \Omega\}$  be the corresponding collection of  $2^q$  probability measures associated with each vector. Also let  $\|P^{\omega'} \wedge P^\omega\|$  denote the affinity between two distributions (i.e.  $\|P^{\omega'} \wedge P^\omega\| = 1 - \sup_A |P^{\omega'}(A) - P^\omega(A)|$ , where the supremum is over all measurable sets), and  $\rho(\cdot, \cdot)$  denotes the Hamming distance between two binary vectors. For any semi-distance  $d$

$$\begin{aligned}
\inf_{\omega} \max_{\omega' \in \Omega} \mathbb{E}_\omega[d^2(f^\omega, f^{\omega'})] &\geq \frac{q}{8} \left( \min_{\omega, \omega': \rho(\omega, \omega') \neq 0} \frac{d^2(f^\omega, f^{\omega'})}{\rho(\omega, \omega')} \right) \\
&\times \left( \min_{\omega, \omega': \rho(\omega, \omega') = 1} \|P^\omega \wedge P^{\omega'}\| \right)
\end{aligned}$$

We now prove Theorem 2.

**Proof. Construction:**

Let  $l = \lfloor c_0 n^{1/(d-1)} \rfloor$  with  $c_0 > 1$  a constant,  $q = l^{d-1}$ ,  $\Omega = \{0,1\}^q$  and  $\epsilon = \frac{1}{l+2}$ . For  $i \in \{1, \dots, l\}$ , let  $a_i = \frac{i+0.5}{l+2}$ . For  $\vec{i} \in \{1, \dots, l\}^{d-1}$ , let  $v_{\vec{i}} = (a_{i_1}, \dots, a_{i_{d-1}})$ . Define  $g : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$  as  $g(\tilde{x}) =$

$$\begin{cases} r + \sqrt{\left(\frac{1}{2} - r\right)^2 - \|\tilde{x}\|_2^2} & \text{for } \|\tilde{x}\|_2 < \frac{1}{2} - r \\ r - \sqrt{r^2 - \left(\frac{1}{2} - \|\tilde{x}\|_2\right)^2} & \text{for } \frac{1}{2} - r \leq \|\tilde{x}\|_2 < \frac{1}{2} \\ 0 & \text{o.w.} \end{cases}$$

for  $\tilde{x} \in \mathbb{R}^{d-1}$ , where  $r \in (0, 1/4)$ , to be specified later. Let  $B = \{(\tilde{x}, x_d) \in [-0.5, 0.5]^{d-1} \times [0, 1] : x_d \leq g(\tilde{x})\}$ . For  $\vec{i} \in \{1, \dots, l\}^{d-1}$ , let  $\underline{B}_{\vec{i}} = \{(\tilde{x}, x_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : ((\tilde{x} - v_{\vec{i}})/\epsilon, x_d - 1/8) \in B\}$  and  $\overline{B}_{\vec{i}} = \{(\tilde{x}, x_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : ((\tilde{x} - v_{\vec{i}})/\epsilon, x_d - (1/8 + r)) \in B\}$ . Let  $\underline{S} = \{x \in \mathbb{R}^d : \exists x' = (\tilde{x}', x'_d) \in [\epsilon, 1 - \epsilon]^{d-1} \times [\epsilon, \frac{1}{8} - \epsilon] \text{ s.t. } \|x - x'\|_2 \leq \epsilon\}$  and  $\overline{S} = \{x \in \mathbb{R}^d : \exists x' = (\tilde{x}', x'_d) \in [\epsilon, 1 - \epsilon]^{d-1} \times [\frac{1}{8} + r + \epsilon, 1 - \epsilon] \text{ s.t. } \|x - x'\|_2 \leq \epsilon\}$ . For any  $\Gamma \subseteq \{1, \dots, l\}^{d-1}$ , let  $\underline{S}_\Gamma = \underline{S} \cup \left( \bigcup_{\vec{i} \in \Gamma} \underline{B}_{\vec{i}} \right)$

and  $\overline{S}_\Gamma = \overline{S} \setminus \left( \bigcup_{\vec{i} \in \Gamma} \overline{B}_{\vec{i}} \right)$ . Let  $\vec{\Gamma}$  be an arbitrary ordering of  $\{1, \dots, l\}^{d-1}$ . Given  $\omega \in \Omega$ , let  $\Gamma(\omega) = \{\vec{\Gamma}_i : \omega_i = 1\}$ , and let  $\underline{S}^\omega = \underline{S}_{\Gamma(\omega)}$ ,  $\overline{S}^\omega = \overline{S}_{\Gamma(\omega)}$ , and  $S^\omega = \underline{S}^\omega \cup \overline{S}^\omega$ .

Let  $p^\omega(x) = \frac{I_{S^\omega}(x)}{\text{Leb}(S^\omega)}$ ,  $f^\omega(x) = MI_{\underline{S}^\omega}(x) - MI_{\overline{S}^\omega}(x)$ , and  $p(y|x)^\omega = \delta(y - f^\omega(x))$ , where  $\delta(\cdot)$  is the Dirac delta (we could also use a conditional distribution that is absolutely continuous with respect to Lebesgue measure; the result would be the same). Finally, let  $P^\omega$  denote the measure on  $\mathbb{R}^{d+1}$  defined by  $p^\omega(x)$  and  $p^\omega(y|x)$ , and  $P_n^\omega$  the corresponding product measure.

**Proof of  $\Omega(1)$  rate:**

Note that  $\text{Leb}(\underline{B}_{\vec{i}}) = \text{Leb}(\overline{B}_{\vec{i}})$ , and so for any  $\omega, \omega'$ ,  $\text{Leb}(S^\omega) = \text{Leb}(S^{\omega'}) = \text{Leb}(\underline{S}) + \text{Leb}(\overline{S})$ . Let  $\lambda = 1/(\text{Leb}(\underline{S}) + \text{Leb}(\overline{S}))$ , i.e.  $\lambda = 1/\text{Leb}(S^\omega)$  for any  $\omega$ .

Let  $\omega, \omega' \in \Omega$  such that  $\rho(\omega, \omega') = 1$  (where  $\rho$  denotes the hamming distance), and WLOG assume  $\omega_i = 0$  and  $\omega'_i = 1$ . Also denote  $\vec{i} = \vec{\Gamma}_i$ . Then the L1 distance

between  $P^\omega$  and  $P^{\omega'}$  is

$$\begin{aligned}
d_1(P^\omega, P^{\omega'}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} |p^\omega(x)p^\omega(y|x) - p^{\omega'}(x)p^{\omega'}(y|x)| dy dx \\
&= \int_{\underline{S}^\omega \cup \overline{S}^{\omega'}} \int_{\mathbb{R}} |\lambda p^\omega(y|x) - \lambda p^{\omega'}(y|x)| dy dx \\
&+ \int_{\overline{B}_i \setminus \underline{B}_i} \int_{\mathbb{R}} \lambda p^\omega(y|x) dy dx + \int_{\underline{B}_i \setminus \overline{B}_i} \int_{\mathbb{R}} \lambda p^{\omega'}(y|x) dy dx \\
&+ \int_{\underline{B}_i \cap \overline{B}_i} \int_{\mathbb{R}} |\lambda p^\omega(y|x) - \lambda p^{\omega'}(y|x)| dy dx \\
&= 0 + \lambda \text{Leb}(\overline{B}_i \setminus \underline{B}_i) + \lambda \text{Leb}(\underline{B}_i \setminus \overline{B}_i) \\
&+ 2\lambda \text{Leb}(\underline{B}_i \cap \overline{B}_i) \\
&= \lambda(\text{Leb}(\underline{B}_i) + \text{Leb}(\overline{B}_i)) \\
&= 2\lambda \epsilon^{d-1} \text{Leb}(B)
\end{aligned}$$

where in the first step we have used the fact that  $x \notin S^\omega \cup S^{\omega'} \Rightarrow p^\omega(x) = p^{\omega'}(x) = 0$ , and divided  $S^\omega \cup S^{\omega'}$  into four non-intersecting components. Then we can bound the affinity of the product measures  $P_n^\omega$  and  $P_n^{\omega'}$  for  $\rho(\omega, \omega') = 1$  as

$$\begin{aligned}
\|P_n^\omega \wedge P_n^{\omega'}\| &\geq (1 - d_1(P^\omega, P^{\omega'})/2)^n \\
&= (1 - \lambda \epsilon^{d-1} \text{Leb}(B))^n.
\end{aligned}$$

For any  $\omega \neq \omega'$ , denoting as  $\omega \wedge \omega'$  the logical and of  $\omega$  and  $\omega'$ , we have, for arbitrary  $\vec{j} \in \{1, \dots, l\}^{d-1}$ ,

$$\begin{aligned}
d^2(f^\omega, f^{\omega'}) &= \sum_{\vec{i} \in \Gamma(\omega \wedge \omega')} \int_{\underline{B}_{\vec{j}} \Delta \overline{B}_{\vec{i}}} M^2 dx + \int_{\underline{B}_{\vec{j}} \cap \overline{B}_{\vec{i}}} 4M^2 dx \\
&= \rho(\omega, \omega') (M^2 \text{Leb}(\underline{B}_{\vec{j}} \Delta \overline{B}_{\vec{j}}) + 4M^2 \text{Leb}(\underline{B}_{\vec{j}} \cap \overline{B}_{\vec{j}})) \\
&= 2\rho(\omega, \omega') M^2 (\text{Leb}(\underline{B}_{\vec{j}}) + \text{Leb}(\underline{B}_{\vec{j}} \cap \overline{B}_{\vec{j}})) \\
&= 2\rho(\omega, \omega') M^2 \epsilon^{d-1} (\text{Leb}(B) + \text{Leb}(B_r))
\end{aligned}$$

where we define  $B_r = \{x \in B : x - (0, \dots, 0, r) \in B\}$ . Then by Theorem 6,

$$\begin{aligned}
&\inf_{\hat{f}} \max_{\omega \in \Omega} \mathbb{E}_\omega[d^2(f^\omega, f^{\hat{\omega}})] \\
&\geq \frac{M^2(l\epsilon)^{d-1}}{4} (\text{Leb}(B) + \text{Leb}(B_r))(1 - \lambda \epsilon^{d-1} \text{Leb}(B))^n.
\end{aligned}$$

10

Also we have

$$\begin{aligned}
\frac{1}{\lambda} \int (f^\omega(x) - f^{\omega'}(x))^2 p^\omega(x) dx &= \int_{S^\omega} (f^\omega(x) - f^{\omega'}(x))^2 dx \\
&= \int (f^\omega(x) - f^{\omega'}(x))^2 dx - \int_{S^{\omega'} \setminus S^\omega} (f^\omega(x) - f^{\omega'}(x))^2 dx \\
&= d^2(f^\omega, f^{\omega'}) - M^2 \text{Vol}(S^{\omega'} \setminus S^\omega) \\
&\geq d^2(f^\omega, f^{\omega'}) - M^2 q \epsilon^{d-1} \text{Leb}(B \setminus B_r) \\
&= d^2(f^\omega, f^{\omega'}) - M^2 \left(\frac{l}{l+2}\right)^{d-1} (\text{Leb}(B) - \text{Leb}(B_r)).
\end{aligned}$$

Since  $\lambda > 1$ ,

$$\begin{aligned}
&\inf_{\hat{f}} \sup_{(p, f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_n \int (\hat{f}(x) - f(x))^2 dP(x) \\
&\geq \inf_{\hat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega \int (f^{\hat{\omega}}(x) - f^\omega(x))^2 p^\omega(x) dx \\
&\geq \frac{M^2(l\epsilon)^{d-1}}{4} (\text{Leb}(B) + \text{Leb}(B_r))(1 - \lambda \epsilon^{d-1} \text{Leb}(B))^n \\
&- M^2 \left(\frac{l}{l+2}\right)^{d-1} (\text{Leb}(B) - \text{Leb}(B_r)).
\end{aligned}$$

Assume  $n \geq 2^d$ . Then  $l \geq 2$  and

$$\left(\frac{l}{l+2}\right)^{d-1} \geq \frac{1}{2^{d-1}}.$$

Clearly  $\text{Leb}(B) \leq \frac{1}{2}$ . Let  $c_0 \geq 3$ . Then  $\epsilon \leq 1/8$  and  $\lambda \leq (1 - 2\epsilon)^{-(d-1)}(1 - 4\epsilon - r)^{-1} \leq 2^{d+1}$ , so

$$(1 - \lambda \epsilon^{d-1} \text{Leb}(B))^n \geq \left(1 - \frac{2^d}{c_0^{d-1} n}\right)^n \rightarrow e^{-2^d/c_0^{d-1}}.$$

So if we let  $c_0 > (2^d/\log(5/4))^{1/(d-1)}$ , then  $e^{-2^d/c_0^{d-1}} > 4/5$  and for sufficiently large  $n$  we will have  $(1 - \lambda \epsilon^{d-1} \text{Leb}(B))^n \geq 4/5$ . Hence,

$$\begin{aligned}
&\inf_{\hat{f}} \sup_{(p, f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_n \int (\hat{f}(x) - f(x))^2 dP(x) \\
&\geq \frac{M^2}{5 \cdot 2^{d-2}} (\text{Leb}(B_r) - 2\text{Leb}(B \setminus B_r)).
\end{aligned}$$

Since

$$\text{Leb}(B_r) = \frac{1}{2} \left(\frac{1}{2} - r\right)^d \frac{\pi^{d/2}}{\Gamma(d/2 + 1)}$$

and

$$\text{Leb}(B \setminus B_r) \leq r \frac{\pi^{(d-1)/2}}{2^{d-1} \Gamma((d-1)/2 + 1)}$$

where  $\Gamma$  is the gamma function, then

$$\begin{aligned} & \text{Leb}(B_r) - 2 \text{Leb}(B \setminus B_r) \\ & \geq \frac{1}{2} \left( \frac{1}{2} - r \right)^d \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} - r \frac{\pi^{(d-1)/2}}{2^{d-2} \Gamma((d-1)/2 + 1)} \\ & \geq \frac{\pi^{d/2}}{2^d \Gamma(\frac{d+1}{2}) d} \left[ (1-2r)^d - \frac{4d}{\sqrt{\pi}} r \right]. \end{aligned}$$

Now let  $r$  be such that

$$(1-2r)^d - \frac{4d}{\sqrt{\pi}} r = \frac{1}{2}$$

(it is easy to see that this can be satisfied by some  $r \in (0, 1/4)$  for any  $d \geq 1$ ). So we have

$$\begin{aligned} & \inf_{\hat{f}} \sup_{(p,f) \in \mathcal{P}_{XY}(\alpha)} \mathbb{E}_n \int (\hat{f}(x) - f(x))^2 dP(x) \\ & \geq \frac{M^2 \pi^{d/2}}{5 \cdot 2^{2d-1} \Gamma(\frac{d+1}{2}) d}. \end{aligned}$$

#### Verifying condition number:

Let  $\tau(A)$  be the condition number of a set  $A$ . Then for arbitrary  $\omega$ ,

$$\tau(S^\omega) = \min \left\{ \tau(\underline{S}^\omega), \tau(\overline{S}^\omega), \frac{1}{2} \inf_{u \in \underline{S}^\omega} \inf_{v \in \overline{S}^\omega} \|u - v\|_2 \right\}.$$

Due to the shape of the function  $g$ , for arbitrary  $\vec{i} \in \{1, \dots, l\}^{d-1}$  we have

$$\tau(\underline{S}^\omega) \geq \min \{ \tau(\partial \underline{S}), \tau(\partial \underline{B}_{\vec{i}} \setminus \partial \underline{S}) \}$$

By definition of  $\underline{S}$  it is easy to see that  $\tau(\partial \underline{S}) = \epsilon$ . Also

$$\begin{aligned} & \tau(\partial \underline{B}_{\vec{i}} \setminus \partial \underline{S}) \\ & = \tau(\{(\tilde{x}, x_d) \in [-\epsilon/2, \epsilon/2]^{d-1} \times [0, 1] : x_d = g(\tilde{x}/\epsilon)\}) \\ & = \epsilon \tau(\{(\tilde{x}, x_d) \in [-1/2, 1/2]^{d-1} \times [0, 1] : x_d = g(\tilde{x})\}) \\ & = \epsilon r. \end{aligned}$$

Since  $r < 1$ , we have  $\tau(\underline{S}^\omega) \geq r\epsilon$ , and similarly  $\tau(\overline{S}^\omega) \geq r\epsilon$ . Now,

$$\begin{aligned} & \frac{1}{2} \inf_{u \in \underline{S}^\omega} \inf_{v \in \overline{S}^\omega} \|u - v\|_2 \\ & \geq \frac{\epsilon}{2} \inf_{u, v \in [-0.5, 0.5]^{d-1}} \|(u, g(u)) - (v, g(v) + r)\|_2 \\ & = \frac{\epsilon}{2} \left( \sqrt{\frac{1}{4} + r^2} - \frac{1}{2} \right) \end{aligned}$$

which is smaller than  $\epsilon r$ , so for  $n$  sufficiently large,

$$\begin{aligned} \tau(S^\omega) & \geq \frac{\epsilon}{2} \left( \sqrt{\frac{1}{4} + r^2} - \frac{1}{2} \right) \\ & \geq \frac{1}{2(c_0 n^{1/(d-1)} + 2)} \left( \sqrt{\frac{1}{4} + r^2} - \frac{1}{2} \right) \\ & \geq n^{-\frac{1}{d-1}} \frac{1}{2(c_0 + 1)} \left( \sqrt{\frac{1}{4} + r^2} - \frac{1}{2} \right) \end{aligned}$$

which completes the proof.  $\square$

**Proof of Theorem 3** First, we derive a general concentration of  $\hat{\mathcal{E}}(f)$  around  $\mathcal{E}(f)$  where  $\hat{\mathcal{E}}(f) = \hat{R}(f) - \hat{R}(f^*) = -\frac{1}{n} \sum_{i=1}^n U_i$ , and  $U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2$ .

If the variables  $U_i$  satisfy the following moment condition:

$$\mathbb{E}[|U_i - \mathbb{E}[U_i]|^k] \leq \frac{\text{var}(U_i)}{2} k! r^{k-2}$$

for some  $r > 0$ , then the Craig-Bernstein (CB) inequality (Craig 1933) states that with probability  $> 1 - \delta$ ,

$$\frac{1}{n} \sum_{i=1}^n (U_i - \mathbb{E}[U_i]) \leq \frac{\log(1/\delta)}{nt} + \frac{t \text{var}(U_i)}{2(1-c)}$$

for  $0 \leq tr \leq c < 1$ . The moment conditions are satisfied by bounded random variables as well as Gaussian random variables (see e.g. Haupt and Nowak [2006]).

To apply this inequality, we first show that  $\text{var}(U_i) \leq 4(M^2 + \sigma^2)\mathcal{E}(f)$  since  $Y_i = f(X_i) + \epsilon_i$  with  $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ . Also, we assume that  $|f(x)|, |\hat{f}(x)| \leq M$ , where  $M > 0$  is a constant.

$$\begin{aligned} \text{var}(U_i) & \leq \mathbb{E}[U_i^2] \\ & = \mathbb{E}[(-(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2)^2] \\ & = \mathbb{E}[(-(f^*(X_i) + \epsilon_i - f(X_i))^2 + (\epsilon_i)^2)^2] \\ & = \mathbb{E}[(-(f^*(X_i) - f(X_i))^2 - 2\epsilon_i(f^*(X_i) - f(X_i)))^2] \\ & \leq 4M^2\mathcal{E}(f) + 4\sigma^2\mathcal{E}(f) = 4(M^2 + \sigma^2)\mathcal{E}(f) \end{aligned}$$

Therefore using CB inequality we get, with probability  $> 1 - \delta$ ,

$$\mathcal{E}(f) - \hat{\mathcal{E}}(f) \leq \frac{\log(1/\delta)}{nt} + \frac{t 2(M^2 + \sigma^2)\mathcal{E}(f)}{(1-c)}$$

Now set  $c = tr = 8t(M^2 + \sigma^2)/15$  and let  $t < 15/(38(M^2 + \sigma^2))$ . With this choice,  $c < 1$  and define

$$a = \frac{t 2(M^2 + \sigma^2)}{(1-c)} < 1.$$

Then, using  $a$  and rearranging terms, with probability  $> 1 - \delta$ ,

$$(1 - a)\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \leq \frac{\log(1/\delta)}{nt}$$

where  $t < 15/(38(M^2 + \sigma^2))$ .

Then, using the previous concentration result, and taking union bound over all  $f \in \mathcal{F}$ , we have with probability  $> 1 - \delta$ ,

$$\mathcal{E}(f) \leq \frac{1}{1 - a} \left[ \widehat{\mathcal{E}}^V(f) + \frac{\log(|\mathcal{F}|/\delta)}{nt} \right].$$

Now consider

$$\begin{aligned} \mathcal{E}(\widehat{f}_{\widehat{\alpha}, \widehat{h}}) &= R(\widehat{f}_{\widehat{\alpha}, \widehat{h}}) - R(f^*) \\ &\leq \frac{1}{1 - a} \left[ \widehat{R}^V(\widehat{f}_{\widehat{\alpha}, \widehat{h}}) - \widehat{R}^V(f^*) + \frac{\log(|\mathcal{F}|/\delta)}{nt} \right] \\ &\leq \frac{1}{1 - a} \left[ \widehat{R}^V(f) - \widehat{R}^V(f^*) + \frac{\log(|\mathcal{F}|/\delta)}{nt} \right] \end{aligned}$$

Taking expectation with respect to validation dataset,

$$\begin{aligned} \mathbb{E}_V[\mathcal{E}(\widehat{f}_{\widehat{\alpha}, \widehat{h}})] &\leq \frac{1}{1 - a} \left[ R(f) - R(f^*) + \frac{\log(|\mathcal{F}|/\delta)}{nt} \right] \\ &\quad + 4\delta M^2. \end{aligned}$$

Now taking expectation with respect to training dataset,

$$\begin{aligned} \mathbb{E}_{TV}[\mathcal{E}(\widehat{f}_{\widehat{\alpha}, \widehat{h}})] &\leq \frac{1}{1 - a} [\mathbb{E}_T[R(f) - R(f^*)] \\ &\quad + \frac{\log(|\mathcal{F}|/\delta)}{nt}] + 4\delta M^2. \end{aligned}$$

Since this holds for all  $f \in \mathcal{F}$ , we get:

$$\begin{aligned} \mathbb{E}_{TV}[\mathcal{E}(\widehat{f}_{\widehat{\alpha}, \widehat{h}})] &\leq \frac{1}{1 - a} \left[ \min_{f \in \mathcal{F}} \mathbb{E}_T[\mathcal{E}(f)] + \frac{\log(|\mathcal{F}|/\delta)}{nt} \right] \\ &\quad + 4\delta M^2. \end{aligned}$$

The result follows since  $\mathcal{F} = \{\widehat{f}_{\alpha, h}^T\}_{\alpha \in \mathcal{A}, h \in \mathcal{H}}$  and  $|\mathcal{F}| = |\mathcal{A}||\mathcal{H}|$ .

□